

**EFFICIENT DEPICTION OF VIDEO FOR SEMANTIC
RETRIEVAL APPLICATIONS BY DIMENSIONALITY
REDUCTION OF VISUAL FEATURE SPACE**

Amarakoon Mudiyansele Randitha Ravimal Bandara

(128029E)

Degree of Doctor of Philosophy

Department of Information Technology

University of Moratuwa

Sri Lanka

March 2021

**EFFICIENT DEPICTION OF VIDEO FOR SEMANTIC
RETRIEVAL APPLICATIONS BY DIMENSIONALITY
REDUCTION OF VISUAL FEATURE SPACE**

Amarakoon Mudiyansele Randitha Ravimal Bandara

(128029E)

Thesis submitted in partial fulfilment of the requirements for the degree
Doctor of Philosophy

Department of Information Technology

University of Moratuwa

Sri Lanka

March 2021

DECLARATION

I declare that this is my own work, and this thesis does not incorporate, without acknowledgment, any material previously submitted for a Degree or Diploma in any other University or institute of higher learning, and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgment is made in the text.

Similarly, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis, in whole or in part in print, electronic, or other media. I retain the right to use this content in whole or part in future works (such as articles or books).

UOM Verified Signature

Signature:

Date: 2/3/2021

The above candidate has carried out research for the PhD thesis under my supervision.

Name of the supervisor I: Dr. Lochandaka Ranathunga

Signature of the supervisor I: ***UOM Verified Signature*** Date: 2/3/2021

Name of the supervisor II: Associate Professor Dr. Nor Aniza Abdullah

Signature of the supervisor II: ***UOM Verified Signature*** Date: 2/3/2021

ABSTRACT

The retrieval of temporal digital visual data, either by a text or visual query, requires automatic interpretation, which includes high-level annotation by object detection and recognition for text query-based retrieval and low-level abstraction for visual query-based retrieval. Both the accuracy and the speed of the interpretation become crucial factors in real-world applications, due to the high density of visual data. This study has focused on reducing the complexity of visual data efficiently by dimensionality reduction techniques for the detection and recognition of objects in videos for both textual annotation and visual query-based video frame retrieval. The contribution of the study includes three approaches, i.e., a novel visual feature descriptor based on colour dithering – namely Salient Dither Pattern Feature (SDPF), novel object segmentation method based on the proposed feature descriptor – namely Refining Superpixel and Histogram of oriented optical flow Clustering (RSHC) –, and a novel self-supervised local descriptor – namely Network-in-Network with Restricted Boltzmann Machine (NIN-RBM). The experimental results make it evident that the SDPF is rotation and scale invariant and computationally efficient yet shows similar object recognition accuracy to the state-of-the-art methods with minimum supervision. The results further revealed that RSHC has successfully utilized SDPF for accurately segmenting individual objects by using a very shallow history of motion. Furthermore, according to the results, NIN-RBM has shown the state-of-the-art correspondence matching performance over the existing deep-learned self-supervised binary descriptors, keeping the computation time at the minimum. The overall results support the conclusions that RSHC is capable of accurately segment objects in a video, and then SDPF can be successfully used for recognizing the segmented objects. Moreover, NIN-RBM can be used to reliably and rapidly retrieve video frames related to any visual query. Since NIN-RBM is a local descriptor, it can be further used for locating of high-level objects and estimating their poses precisely, to improve the details of semantics retrieved from video data.

Keywords: dimensionality reduction, colour dithering, deep learning, video segmentation, object recognition, correspondence matching, binary descriptor

ACKNOWLEDGEMENT

First and foremost, I consider it is my bounden duty to record here my sincerest gratitude and appreciation to my supervisor, Dr. Lochandaka Ranathunga for his productive cooperation, guidance, and supervision extended throughout this research project. If not for his encouragement and support, this project would never have been a reality. Similarly, I unreservedly thank my co-supervisor, Associate Professor Dr. N.A. Abdullah from the University of Malaya, Malaysia for guiding me to succeed in this endeavour.

I express my heartfelt thanks to the Vice-Chancellor of the University of Moratuwa, the Dean of the Faculty of Information Technology and the Head of the Department of Information Technology, the non-academic staff members of the Faculty of Information Technology, University of Moratuwa for the opportunity given to commence my research work at the University of Moratuwa and facilitating me to carry out the same successfully. Further, I extend my sincere gratefulness to the Vice-Chancellor of the University of Sri Jayewardenepura, the Dean of the Faculty of Applied Sciences and the Head of the Department of Computer Science, and my colleagues in the academic staff for their assistance in many ways, which was a blessing for completing this research study. Thanks, are also due to the chairman and the staff of the National Research Council, Sri Lanka for granting a research scholarship under grant number NRC/12-017 to financially support this endeavour. Also, I must be thankful to the Senate Research Committee of University of Moratuwa for the financial support provided to carry out this study. Furthermore, I am indebted to the CEO of LK Domain Registry for granting me the Prof. V.K. Samaranayake Research Grant in order to continue my research studies at the University of Malaya, Malaysia.

I declare my salutation and admiration for all the esteemed authors, researchers, and philosophers for their great theories, research, publications, and ideas, which have been an enormous support to enhance this research work.

The support and motivation provided by my postgraduate friends, Mr. K.A.S.H. Kulathilake, Mr. V. Senthoran, Mr. B. Hettige and Mrs. M. Sirisuriya who are

affiliated with the Faculty of Information Technology, University of Moratuwa, too deserve mentioning with a debt of gratitude.

I am grateful to my parents, my sister, and brother, for all their encouragement and support extended right throughout this endeavour. If not for them, this project might not have been a reality. Last but not least, it is the dedication of my loving wife and daughter who were bearing all the burdens without passing them to me throughout the past few years that made me complete this research. Finally, I am grateful to all those who assisted me in numerous ways during the course of this research.

TABLE OF CONTENTS

Declaration	i
Abstract	ii
Acknowledgement.....	iii
Table of Contents	v
List of Figures	x
List of Tables.....	xiii
List of Abbreviations.....	xiv
Chapter 1 Introduction	1
1.1 Digital video.....	2
1.1.1 Background of digital videos	2
1.1.2 Secondary tools for video applications	2
1.2 Digital video understanding.....	4
1.2.1 Video from the human’s point of view	4
1.2.2 Video from the machine’s point of view	5
1.2.3 Different aspects of machine’s understanding.....	6
1.3 Computer vision.....	7
1.3.1 Image processing	7
1.3.2 Artificial intelligence	8
1.4 Handcrafted vs learned features	9
1.5 General process of video data interpretation	9
1.6 Problem in brief	10
1.6.1 Text query based retrieval.....	12
1.6.2 Visual query based retrieval.....	12
1.6.3 Scope of the study	13
1.7 Study aim and objectives	14

1.8 Organization of the thesis	16
1.9 Summary	17
Chapter 2 Research Background.....	18
2.1 Scene and object recognition	18
2.1.1 Handcrafted feature descriptors	18
2.1.2 Dimensionality reduction.....	20
2.1.3 Deep learned features.....	21
2.2 Video segmentation	22
2.2.1 Semantic segmentation	23
2.2.2 Supervised and unsupervised segmentation.....	23
2.2.3 Data clustering	24
2.3 Known object detection	26
2.3.1 Correspondence matching.....	26
2.4 Binary-valued local feature descriptors	27
2.5 Summary	31
Chapter 3 Low-Dimensional Feature for Object Recognition.....	32
3.1 Overview	32
3.2 Colour dithering as a dimensionality reduction technique	33
3.3 Salient Dither Pattern Feature (SDPF).....	33
3.3.1 SDPF feature point extraction.....	34
3.3.2 SDPF descriptor	37
3.4 Dither Density Descriptor (DDD).....	44
3.5 Improved Hessian based salient dither pattern feature (HSDPF)	46
3.5.1 HSDPF descriptor	53
3.5.2 Classification of SDPF descriptor.....	59
3.6 Summary	59

Chapter 4 Objects Segmentation	60
4.1 Overview of the proposed methods to segmentation	60
4.2 K-means based feature clustering for segmentation	61
4.2.1 Selection of properties	61
4.2.2 Attributes of the data for clustering	63
4.2.3 Estimation of K in K-means	65
4.3 Segmentation of video frames to an unknown number of objects	66
4.3.1 Simple Linear Iterative Clustering	67
4.3.2 Histogram of Oriented Optical Flow	68
4.3.3 Clustering feature points	68
4.3.4 Refining process of superpixels	72
4.4 Summary	74
Chapter 5 Correspondence Matching for Object Detection	75
5.1 Overview	75
5.2 Training NIN model	77
5.3 Training RBM	80
5.4 Calculate representative code	82
5.5 Fine tuning NIN	84
5.6 Summary	85
Chapter 6 Experimental Setup and Validation Methods	87
6.1 Evaluation of object recognition	87
6.1.1 Datasets	87
6.1.2 Evaluation metrics	88
6.1.3 Assessing the invariant properties of SDPF	91
6.2 Evaluation of segmentation	91
6.2.1 Datasets	92

6.2.2 Evaluation metrics	92
6.3 Evaluation of correspondence matching	93
6.3.1 Datasets	93
6.3.2 Evaluation metrics	96
6.4 Summary	98
Chapter 7 Results and Discussion	100
7.1 Invariant properties of SDPF	100
7.1.1 Rotational invariance assessment	100
7.1.2 Scale invariance assessment	103
7.2 Assessing the performance of SDPF and SDPF-DDD	104
7.2.1 Assessing the classification performance of HSDPF	109
7.2.2 Assessment of the cost of HSDPF	113
7.3 Experimental results of object segmentation	117
7.3.1 Segmentation with an estimated number of objects	117
7.3.2 Segmentation without estimating the number of objects	120
7.4 Evaluation of correspondence matching of NIN-RBM	123
7.4.1 Evaluation of instance retrieval	133
7.4.2 Tasks from the HPatches benchmark	137
7.4.3 Evaluation of the cost of computation	140
7.5 Summary	143
Chapter 8 Conclusion and Recommendations	145
References	151
APPENDIX A : Table of pre-calculated error diffusion coefficient vectors	159
APPENDIX B : Analysis of T_K , the fraction of the average dissimilarities	160
APPENDIX C : Nearest neighbour classification of SDPF	161
APPENDIX D : Number of objects vs number of SDPF points	162

APPENDIX E : Performance of HSDPF on segmented objects.....	163
APPENDIX F : Modified Network in Network model.....	164
APPENDIX G : Sample images of inputs and internal response of NIN-RBM.....	165
APPENDIX H : Sample Images from the Datasets	171
APPENDIX I : Publications based on this research study.....	173

LIST OF FIGURES

Figure 1.1: Two major sub-processes used in video retrieval applications, namely (1) text query based retrieval by annotation and indexing, and (2) visual query based retrieval by correspondence matching. The focus of the study is shown inside the blue frame.....	11
Figure 3.1: 3x3 neighbor dither patterns with the anatomy of a single dither pattern.....	35
Figure 3.2: The analogy of the function $F_c(x)$ to hue and intensity value quantization.....	41
Figure 3.3 The set of SDPF points allocated to the set of distance bins. The yellow markers show the SDPF points, whereas the green cross marker is the centroid. The coloured bands are the distance bins.....	41
Figure 3.4: The experimental results to find the optimal dimension for SDPF descriptor.....	43
Figure 3.5: Coding grayscale with dither density utilizing a single bit per pixel. The spatial density is calculated using square-shaped regions.	44
Figure 3.6: Splitting the dither image to several binary images based on the dither colours	45
Figure 3.7: The experimental results of finding the optimal number of colours and circular regions..	46
Figure 3.8: Improved ED-Dithering algorithm preserves the colour contrast. (a) two regions with slightly different colours (b) Linearly quantized (c) Enlarged dither patterns (d) ED colour dithering-based quantization.....	47
Figure 3.9: Dither colour set in RGB space	49
Figure 3.10: Binary search tree of dither colours	49
Figure 3.11: Different permutations of the same set of colours with the resultant overall colours.	50
Figure 3.12: Different instances of an object in the SDPF algorithm. (a) the original image, (b) dithered image (c) extracted SDPF points, (d) calculated dominant orientation.....	56
Figure 3.13: The accuracy vs. the dimension over distance axis k_d and chromatic axis k_c . (a) $k_a = 4$, (b) $k_a = 8$, (c) $k_a = 12$ and (d) $k_a = 15$	58
Figure 4.1: Relative motion difference of feature points detected at different depth with a moving camera. (a) front view (b) top view	62
Figure 5.1: The novel NIN-RBM hybrid model with the proposed learning method	77
Figure 5.2: Newly adapted NIN model by replacing the last pooling layer	78

Figure 7.1: A sample of images that were used to evaluate the invariant properties.....	101
Figure 7.2: Probability of rotated images classified in their true image class, obtained for example image 1.....	101
Figure 7.3: Probability of rotated images classified in their true image class, obtained for example image 2.....	102
Figure 7.4: Average probability of rotated images classified in their true image classes (considering all the images).....	102
Figure 7.5: Resized versions of an input with the scaling factor.....	103
Figure 7.6: Mean probability of predicting an input to its ground truth.....	103
Figure 7.7: Comparison of retrieval precision; (a) ten visual concepts in Corel dataset, (b) six visual concepts in Caltech dataset.	105
Figure 7.8: Comparison of recall; (a) ten visual concepts in Corel dataset, (b) six visual concepts in Caltech dataset.	106
Figure 7.9: Comparison of F1 Scores; (a) ten visual concepts in Corel dataset, (b) six visual concepts in Caltech dataset.	107
Figure 7.10: The performance of recognizing objects; (a) dataset: ALOI-View, (b) dataset: Coil-100 dataset, in terms of class average precision.....	110
Figure 7.11: The performance of recognizing objects in ALOI-ill in terms of class average precision	110
Figure 7.12: Sample images from five object categories in ALOI-ill dataset. Each column contains two images from a single object category, captured under different illumination conditions.	111
Figure 7.13: Performance of recognizing objects that are in different orientation in ALOI-View. (a) without data augmentation (b) with data augmentation	112
Figure 7.14: Comparison of the three algorithms with the sequence 16E5. (a) Completeness measure. (b) Spatial accuracy measure.....	118
Figure 7.15: Clustering SDPF points in a video frame using K-means. (a) source frame (b) optical flow of SDPF (c) clustered SDPF (d) ground truth segments	119
Figure 7.16: Performance of RSHC, K-means-8D and EM over consecutive frames in the sequence 16E5. (a) Completeness measure. (b) Spatial accuracy measure.	121

Figure 7.17: Clustering SDPF points in a video frame using the proposed methods. (a) Source frame. (b) manual segmentation (c) K-means-8D clustered (d) clustered with RSHC.....	123
Figure 7.18: Correct and incorrect output of NIN-RBM for Brown Dataset. First row shows four sample pairs of matched patches from Yosemite, the second one from Notre Dame and the last row of is from liberty subsets.....	129
Figure 7.19: ROC of binary-valued descriptors with all combinations of all cross-category training and testing configurations with Browns dataset.....	131
Figure 7.20 Performance over the three tasks namely retrieval, verification and matching. The coloured circular bullets indicates the performance over the three different challenging levels found in HPatches dataset. binary-valued or real-valued categories are denoted with different background colours.	139
Figure 7.21: Mean time taken for feature encoding. Dataset: CIFAR-10	143

LIST OF TABLES

Table 7.1. Dimension and average extraction time of feature descriptors	104
Table 7.2: Averaged results obtained from the experimental setups of seven visual descriptors with the two datasets	108
Table 7.3 Computational cost of the descriptors	114
Table 7.4. Computational details of HSDPF	116
Table 7.5: The average completeness and the average spatial accuracy error of the segmentation with unknown number of objects	122
Table 7.6. Comparison of patch matching performance of NIN-RBM	125
Table 7.7 Results of correspondence matching task with RomePatches	133
Table 7.8 Performance over instance retrieval task with Holidays, Oxford and Paris in terms of mAP %	136
Table 7.9 Model size and number of parameters of the base models	140

LIST OF ABBREVIATIONS

Acronym	Definition
ALOI	Amsterdam Library of Object Images
ANN	Artificial Neural Network
ARM	Advanced RISC (Reduced Instruction Set Computing) Machine
BOW	Bag of Words
BRIEF	Binary Robust Independent Elementary Features
BRISK	Binary Robust Invariant Scalable Keypoints
CCTV	Closed-Circuit Television
CDPC	Compact Dither Pattern Code
CIELAB	International Commission on Illumination. L, A and B are colour components
CIFAR	Canadian Institute for Advanced Research
CKN	Convolutional Kernel Network
CNN	Convolutional Neural Network
CNNH	Convolutional Neural Network with Hashing Layer
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture
DBD-MQ	Deep Binary Descriptor with Multi-Quantization
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DDD	Dither Density Descriptor
DNN	Deep Neural Network
DSH	Deep Supervised Hashing

Acronym	Definition
EHD	Edge Histogram Descriptor
FAST	Features from Accelerated Segment Test
FC-GPHOG	Fused Colour-Gabor Pyramidal Histogram of Oriented Gradient
FREAK	Fast Retina Key- point
GAP	Global Average Pooling
GBRBM	Gaussian-Bernoulli Restricted Boltzmann Machine
GPHOG	Gabor Pyramidal Histogram of Oriented Gradient
GPU	Graphic Processing Unit
HOG	Histogram of Oriented Gradients
HOOF	Histogram of Oriented Optical Flow
HSDPF	Hessian based Salient Dither Pattern Feature
HSV	Hue Saturation Value
HTD	Homogeneous Texture Descriptor
IFV	Improved Fisher Vector
ILSVRC	ImageNet Large Scale Visual Recognition Challenge
ITQ	Iterative Quantization
LAP	Local Average Pooling
LDA	Linear Discriminant Analysis
LDB	Local Difference Binary
LPP	Locality Preserving Projection
LSH	Local Sensitivity Hashing
MLP	Multi-Layer Perceptron

Acronym	Definition
NIN	Network in Network
NIN-RBM	Network in Network with Restricted Boltzmann Machine
ORB	Oriented FAST and Rotated BRIEF
PCA	Principle Component Analysis
PHOG	Pyramidal Histogram of Oriented Gradient
RAM	Random Access Memory
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machine
ReLU	Rectified Linear Unit
RFD	Receptive Fields Descriptor
RGB	Red, Green, and Blue
ROC	Receiver Operating Characteristics
RPi	Raspberry Pi
RSHC	Refining Superpixels using HOOF and Colour
SDPF	Salient Dither Pattern Feature
SGD	Stochastic Gradient Descent
SIFT	Scale Invariant Feature Transform
SIMD	Single Instruction Multiple Data
SKAR	Smoothed Keypoint-Matching Ratio
SLIC	Simple Linear Iterative Clustering
SURF	Speeded Up Robust Feature
SVM	Support Vector Machine

Acronym	Definition
TBD	Texture Browsing Descriptor
VGG	Visual Geometry Group